大規模な化合物データベースからの類似化合物探索手法

河村 元† 瀬尾 茂人† 竹中 要一† 松田 秀雄†
大阪大学大学院情報科学研究科†

タンパク質，化合物データベース等を用いて創薬のための化合物探索を行う際には，複数の大規模な創薬関連データベースを探索しなければならない．しかしながら，データベースの容量が膨大であるため単純に検索すると組合せ爆発を起こす恐れがある．そこで，本研究では，複数の創薬関連データベースを用いて，ターゲットタンパク質から相互作用する化合物を，組合せ爆発を抑えながら探索する手法を提案する．具体的には，ターゲットタンパク質から相同なタンパク質集合を探索し，相互作用情報と化合物探索手法によって，相互作用する化合物を推定する．本手法を既知のタンパク質・タンパク質－化合物相互作用，化合物のデータベース群に対して評価実験を行い，有用性を示す．

キーワード　化合物，データベース，化合物探索，Random Forest，proximity measure

1．まえがき

生体内で目的とするレセプターへ特異的に結合する化合物（リガンド）を見つけるための一般的な手法は化合物の部分構造に着目し、構造類似度を評価する探索手法である。特に疾患の目標となるタンパク質（ターゲットタンパク質）に作用するリガンド候補を探索する場合にはこのような類似構造化合物を探索する手法は非常に有用である[1]．タンパク質の作用が分かっている場合や天然の有機化合物や合成物が何らかの作用を示すことが既知の場合には化合物データベースからリガンドの候補を探し出すことが可能である[2]．この手法では，新規ターゲットタンパク質へ作用するリガンドの情報が創薬のプロセスの後半で既知でなければならない．そこで新規ターゲットタンパク質に作用するリード活性化合物の探索手法として，

(1) ターゲットタンパク質から BLAST を用いて相同性検索を行う．
(2) 相同なタンパク質に対して作用既知の化合物を見つける．
(3) 見つけた化合物と構造が類似した化合物の探索を行う．

という複数の検索段階を持った探索手法が提案されている[3]．L_1がターゲットタンパク質T_1のリガンドであり，かつL_2はタンパク質T_2のリガンドであるとすると，T_1とT_2が相同であればL_1とL_2の活性も「類似」であると考えることができる．これらL_1とL_2は共に特有の構造を有して

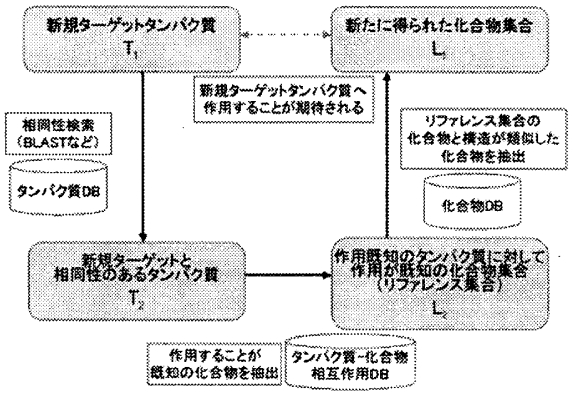

図 1: ターゲットタンパク質からの類似構造に基づく相互作用探索

いると予想できる．このような考え方を利用して活性未知の新規ターゲットタンパク質T_1から相同性検索，相互作用検索、化合物探索を組合せてリガンドL_1を探索できる（図１）．この探索結果のリガンドL_1は新規ターゲットタンパク質T_1に作用する可能性があると考えられる．しかし，化合物を探索するために一般的に利用されているMACCSキーとTanimoto係数を利用した類似化合物の探索手法は，活性クラスに対応した部分構造の重要度を反映できないので特有の構造を持つ化合物を見落してしまう可能性がある．さらに，このようなタンパク質の相同性も考慮したリガンドの探索方式は，いくつもの探索手法を連結するため検索結果が急激に増大し大量の無関係の化合物が拾いあげられる可能性がある．特に，化合物探索ではタンパク質の相同性を反映しながら多クラスの活性を識別し精度良く化合物候補を検索する必要がある．

A Search Method for Similar Compounds in Very Large Compound Database
Gen Kawamura†, Shigeto Seno†, Yoichi Takenaka†, Hideo Matsuda†
†Graduate School of Information Science and Technology, Osaka University

2．提案手法

本研究では，図1の手法のリガンドL_2からリガンドL_1の化合物の探索部分へ，機械学習法のRandom Forestのproximity measure[4] [5]とTanimoto係数[1]による化合物類似尺度を線形判別分析[6]によって組合せたスコアを利用した探索法を使用する．この化合物探索手法は活性化合物のproximity measureとTanimoto係数の正誤の傾向を線形判別法で組合せてスコア化することで精度よく活性化合物の探索を行うことが出来る．本手法をタンパク質の相同性検索や相互作用検索と組合せることで活性未知のターゲットタンパク質からのリード活性化合物の探索が可能である．

3．評価実験

評価実験のデータとしてドーパミン D2 レセプター（以下，D2 レセプター）の(SwissProt ID: P14416) をターゲットタンパク質として想定した．この D2 レセプターに対して相同な D1 レセプター(SwissProt ID: P21728)，D3 レセプター(SwissProt ID: P35462)，D4 レセプター(SwissProt ID: P21917)を考える． D2 レセプターに対して BLAST で相同性検索を行った結果が表 1 である．この表からドーパミン D3，D4，D1 レセプターの順序で相同性が高い(E-value が小さい) ことが分かる．

表 1: ドーパミン D2 レセプターに対する相同性探索

タンパク質	E-value
ドーパミン D3 レセプター	3E-109
ドーパミン D4 レセプター	5E-56
ドーパミン D1 レセプター	2E-29

図 2 のグラフは提案手法によって D1 リガンド，D3 リガンド，D4 リガンドの識別器を構築し．これらリガンドをリファレンスとして D2 リガンドを含むリファレンス化合物を検索した化合物ランキング順位の結果である．上位約 750 位で D2 リガンドが D3 リガンド，D4 リガンド，D1 リガンドの順で選択された．これは D3 レセプターが D2 レセプターの相同性に対応している．D3 リガンドをリファレンスとして D2 リガンドを探索した場合には，本手法では D3 として識別される化合物が多く取得できるということである．ランキングの上位の化合物を取得すれば D2 レセプターに相同性のある順序にしたがって D2 リガンドが取得できることが分かる．このことから本手法は未知タンパク質からの活性化合物探索に有効であることが分かる．

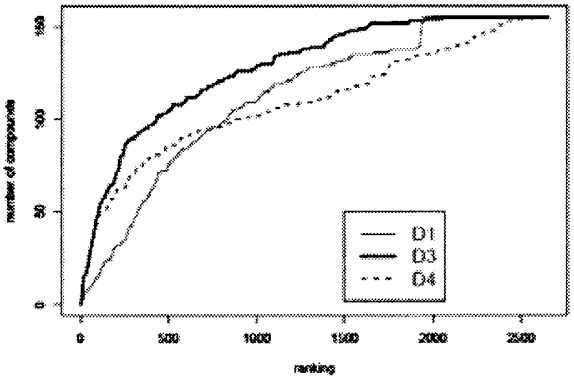

図 2: D1，D3，D4 リガンドをリファレンスにした D2 リガンドのランキング

4．むすび

相同なタンパク質に対して活性既知のリガンドからの類似化合物探索に対して線形判別分析法による化合物類似度のスコア組合せによる探索法を行うことで化合物の活性の多クラス分類も容易に対応付けることができた．相同性の高いタンパク質から対応付いた既知リガンドを元にして提案した方法で活性探索を行うことで，ターゲットタンパク質に作用する化合物候補が精度良く探索できる．

さらに評価の結果から相同性の高い化合物をリファレンスにするほど目的とするタンパク質に対するリガンド候補が正確に検索できるので，あらかじめ既知リガンド数の上限値などのパラメータの調査などを行うことが必要と思われる．

謝辞

本研究は，サイエンスグリッド NAREGI プログラムおよび科学研究補助金特定領域研究「情報爆発」の成果である．

参考文献

[1] J. Gasteiger and T. Engel. *Chemoinformatics*. WILEY-VCH, 2003.

[2] 田沼靖一. ゲノム創薬. 化学同人, 2003.

[3] A. Schuffenhauer, P. Floersheim, and P. Acklin. "Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins", *J. Chem. Inf. Comput. Sci.*, vol. 43, pp. 391–405, 2003.

[4] L. Breiman. "Random forests", *Machine Learning*, vol. 45, pp. 5–32, 2001.

[5] V. Svetnik and A. Liaw. "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling", *J. Chem. Inf. Comput. Sci*, vol. 43, pp. 1947–1958, 2003.

[6] S. Balakrishnama and A. Ganapathiraju. Linear Discriminant Analysis – *A Brief Tutorial,* 1998.